LANGUAGE DOCUMENTATION & CONSERVATION

Vol. 3, No. 1 (June 2009), pp. 87-99
http://nflrc.hawaii.edu/ldc/
http://hdl.handle.net/10125/4425

Data Processing and its Impact on Linguistic Analysis

Anna Margetts

Monash University

The Saliba-Logea documentation project has been working toward a web-based text database with text-audio linkage and searchable annotations. In this article, I discuss the impact that the nature of data processing can have on linguistic analysis, and I demonstrate this on the basis of two research topics: the positioning of Postpositional Phrases and the distribution of plural markers. Saliba-Logea PPs can be ambiguous as to whether they belong to the preceding or following clause. To investigate whether there is a correlation between a PP's position and its semantic role, text-only transcriptions turn out to be insufficient. The second question relates to the Saliba-Logea plural suffix, which originally occurred only on nouns with human referents. However, some speakers use it in novel contexts, and in order to investigate these extended uses and who drives them, access to metadata about the speakers is required. I show that text-audio linkage can be a prerequisite for analyzing syntactic constructions and that access to metadata can have a direct effect on the linguistic analysis.

1. INTRODUCTION. This article is based on research within the Saliba-Logea language documentation project, which has been funded since 2004 by of the Documentation of Endangered Languages (DoBeS) program of the Volkswagen Foundation [1].[1] The DoBeS program stipulates that the primary focus of funded documentation projects is on collecting texts, based on the assumption that grammar and lexicon can, at least to some extent, be derived from these, but not vice versa. Project funding is linked to certain standards of data processing and archiving. Minimally, texts need to be transcribed, text-audio linked and provided with a translation. Subsets of texts should be presented with morpheme break-down, and interlinearization (to facilitate morphological analysis of the rest of the corpus if necessary) and phonetic transcription (so that orthographic conventions of the corpus can be reconstructed).

---

[1] I would like to acknowledge the contribution to this paper by Andrew Margetts, who was behind the technical aspects reported here, including creating the database and manipulating Toolbox to do things that it wouldn't normally do. I also thank Carmen Dawuda for collecting most of the Logea data. An earlier version of this paper was presented at the 2007 DoBeS meeting at the Max Planck Institute in Nijmegen and at the 7th International Conference of Oceanic Linguistics in Noumea, New Caledonia, also in 2007. I would like to thank the audiences at these meetings for feedback and comments. Nick Thieberger had some useful comments on the original conception of this paper. As always, I cordially thank the communities and individuals on Saliba and Logea Island who have supported my research.

E-ISSN 1934-5275

In this article, I discuss the impact that our data processing can have on the linguistic analysis and I demonstrate this on the basis of two research topics: the positioning of different types of PPs within the clause, and the spreading distribution of the plural suffix. The first example illustrates that text-only transcriptions are insufficient for the analysis of certain morpho-syntactic constructions. The second example highlights the importance of searchable metadata for linguistic analysis and for the interpretation of morpho-syntactic change.

1.1 SALIBA-LOGEA LANGUAGE. There are about 2500 speakers of Saliba-Logea, who live predominantly as subsistence farmers on Saliba and Logea Island and the surrounding area in Milne Bay Province of Papua New Guinea. The two main dialects, Saliba and Logea, are mutually intelligible and differ mainly in their lexicon.

The language has SOV and Genitive-Noun word order and postpositions, rather than VO, Noun-Genitive, and prepositions, as found in most Oceanic languages. This pattern has been attributed to historical contact between the Papuan Tip cluster and non-Austronesian languages (c.f. Bradshaw 1982, Ross 1988).

There is nominative–accusative alignment with obligatory subject prefixes and not quite so obligatory object suffixes.[2] The object suffixes cross-reference certain types of object arguments, essentially those that are highly individuated, e.g., by being marked as specific or definite (c.f. Margetts 2008a, 1999). Both nuclear-layer and core-layer serialization are attested in the language (Margetts 2004, 2005).

1.2 DATA FLOW. The data flow in our project normally proceeds as follows:

1. Recording (mainly on mini digital video)
2. Capturing to disk & creating Adobe "Premiere" project file
3. Identifying & cutting sessions
4. Transcoding to mpeg and wave files
5. 'Chunking' text into intonation units & creating audio linkage in "Transcriber"
6. Transcribing text (research assistants in Papua New Guinea)
7. Converting files to "The Linguist's Toolbox" and/or "ELAN" files using the "Transcriber-to-Toolbox Converter" [2]
8. Translating into English
9. Interlinearizing (a subset of the texts)

[2] The subject prefixes are written separately from the verb in the Saliba trial orthograpphy developed by SIL (c.f. Oetzel and Oetzel 1997).

A "session," as referred to in point 3, can be any amount of recorded data that we consider to be one event. There are typically several sessions on one tape (e.g., several stories or interviews), but occasionally one session goes over several tapes (e.g., the filming of a canoe race).

In point 5, our notion of "chunking" refers to associating chunks of the audio recording with empty lines in the "Transcriber" program (essentially text-audio linkage but still without the transcribed text). Files prepared in this way are then sent to native-speaker research assistants, who transcribe the text line by line in Transcriber. This completes the text-audio linkage. Our texts are chunked into intonation units—i.e., units defined by intonation contour. Unit boundaries often coincide with pauses, but this is not necessarily the case. We chose these units rather than, e.g., intonational sentences because the latter are essentially rhetorical units of very variable length, often corresponding to what might be considered a whole thematic paragraph in a written text.

In this workflow, metadata are noted on different levels of data organization. The recording date is normally noted tape by tape; the location of the recording is noted session by session. The identity of the speaker is marked for each intonation unit (rather than, for example, for the entire session). This is important, as it allows us to relate each individual intonation unit to relevant speaker metadata. The lower units of organization inherit the information of the higher units so that, for example, intonation units inherit their metadata about the recording date and location from the tape and the session respectively.

The dataset used for this paper consists of texts recorded between 1995 and 2007 and includes all texts that were transcribed and part of the corpus by May 2007. This dataset comprised eighty-seven texts of about 14850 intonation units, from sixty-two speakers between the ages of about ten to eighty.

2. EXAMPLE ONE: POSITIONING OF PPS. As mentioned, Saliba-Logea has postpositions rather than prepositions. Most common is the general postposition *unai* (and its plural form *udiyedi*), which can mark a range of different semantic roles including Location, Goal, Source, Addressee, Recipient, Cause, and Instrument. Examples where *unai* introduces these roles are given in (1) to (7).[3]

(1) *Nige* *yo-di* *numa* *unai* *se* *keno-keno*.
NEG CLF-3PL.POSS house PP.SG 3PL RED-sleep
'They didn't sleep in their houses.' Beyabeyana_02CZ_1124

(2) *Meimeilahi* *kabo* *ku* *laoma* *yo-gu* *numa* *ne* *unai*.
Afternoon TAM 2SG come CLF-1SG.POSS house DET PP.SG
'Come to my house in the afternoon.' TBlaki_05AC_0066

[3] In addition to those listed in the Leipzig Glossing Rules the following abbreviations are used: ASSOC, associative; ANA, anaphoric marker; CONJ, conjunction; PP, postposition; RED, reduplication.

(3) *Babadao* *unai* *ye* *laoma*.
ancestors PP.SG 3SG come
'It comes from our ancestors.' Beyabeyana_02CZ_0348

(4) *Ya* *hedede* *lao* *unai* *na* *ye* *hetubu*.
1SG tell go PP.SG and 3SG begin
'I'll tell her to start.' BasketWeaving_05AA_0121

(5) *Ye* *mosei* *ka-na* *kaha* *wa* *unai*.
3SG give CLF-3SG.POSS sibling ANA pp.sg
'He gave it to his friend.' Abs-Rel_02DO_0015

(6) *Nige* *se* *kita-kita-di* *unai* *se* *matausi*.
neg 3pl RED-see-3PL PP.SG 3PL afraid
'They had never seen them that's why they were frightened.' Torres_01AC_0277

(7) *Taba* *nige* *weku* *unai* *ta* *tutu*.
IRR NEG stone PP.SG 1INCL pound
'We don't pound it with a stone.' Garden_02CY_0127

The position of the PP within the clause is not fixed; it can occur before or after the verb as in examples (8) and (9).

(8)

Figure		Verb		Ground		PostP
Simai	*wa*	*ye*	*tu-tuli*	*leiyaha*	*ne*	*unai*.
cat	ANA	3SG	RED-sit	mat	DET	PP.SG

'The cat is sitting on the mat.'

(9)

Figure		Ground		PostP	Verb	
Simai	*wa*	*leiyaha*	*ne*	*unai*	*ye*	*tu-tuli*.
cat	ANA	mat	DET	PP.SG	3SG	RED-sit

'The cat is sitting on the mat.'

From this a number of research questions arise. Questions that need to be investigated include whether PPs more commonly precede or follow the verb; and whether the position of the PP varies according to the role being introduced (e.g., goal vs. location or source). In the remainder of this section, I will demonstrate how the treatment of our data and the functionality of our database can affect the investigation of these research questions.

2. 1 THE PROBLEM OF CLAUSE BOUNDARIES. A serious problem in answering the research questions above is that clause boundaries are not always overt in Saliba-Logea. Where clause boundaries are overtly marked, for example, by a conjunction or particle before or after the PP, as in (10) to (12) or an interjection as in (13), the question of whether a PP belongs to the preceding or the following clause is easy enough to answer.

(10)

Ka	*bahe*	***KABO***	*bosa*	*wa*	*unai*	*ka*	*usa*.
1EXCL	carry	and.then	basket	ANA	PP.SG	1EXCL	insert

'We carry (it) **AND THEN** we put (it) in the basket.' Gulewa_01AH_086

(11)

Ka	*kala*	*NA*	*KABO*	*unai*	*ka*	*keno*.
1EXCL	spread	CONJ	and.then	PP.SG	1EXCL	sleep

'We spread it out **AND** sleep on it.' Nogi_01AQ_095

(12)

Ye	*tole-sae*	*taibolo*	*wa*	*kewa*	*wa*	*unai*	***NA***	*ye*	*kabi-sae*.
3SG	put-up	table	ANA	top	ANA	PP.SG	CONJ	3SG	reach-up

'He put it up on top of the table **AND** he reached up.' PairedFilms_01AW_0201-2

(13)

Ye	*dobi*	***EEE***	*isutete-na*	*wa*	*unai*	*ye*	*dui*.
3SG	go.down	DUR	point-3SG.POSS	ANA	PP.SG	3SG	bathe

'She went down and down **AND** at the point she bathed.' Gagagageniyo_01CP_059

However, there are cases where sequences of VERB PP VERB appear in the text without any particles or conjunctions separating the PP from either of the two clauses. In such cases it can be unclear whether the PP belongs to the first or the second clause. This constitutes an obvious problem if we want to investigate whether the PP typically appears at the beginning or the end of the clause. And since *unai* can introduce such a range of roles there can be some ambiguity about the meaning that is expressed. Consider the examples in (14) to (18), where the translations in (a) and in (b) both seem possible:

(14)

Se	*bui-gabae*	*se*	*dobima*	*Rossel*	*unai*	*se*	*laoma*	*snakepasis*
3PL	go.down	3PL	come.down	Rossel	PP.SG	3PL	come	snake.passage

(a) 'They turned away, [went down to Rossel Island] and came to the Snake Passage.'
(b) 'They turned away, [went down] and at Rossel Island they came to the Snake Passage.' Torres_01AC_0088-89

(15) *Se* *lao-ma* *numa* *wa* *unai* *se* *gwau*.
3PL go-hither house ANA PP.SG 3PL heap
(a) '[They come to the house] and gather (things).'
(b) 'They come and [gather (things) in the house].' Giyahi_01AA_073

(16) *Ye* *lao* *unai* *ye* *keno-keno*.
3SG go PP.SG 3SG RED-sleep
(a) '[He went there] and was resting.'
(b) 'He went and [was resting there].' TBLaki_01AG_221

(17) *Se* *lao* *nukula* *unai* *maina* *se* *boli* *ye* *gehe*.
3PL go jungle PP.SG string 3PL cut 3SG finish
(a) '[They went to the bush] and cut string.'
(b) 'They went and [in the bush they cut string].' MakingSagoRoof_01AW_016

(18) *Ye* *dobi* *unai* *ye* *yoli*.
3SG go.down PP.SG 3SG sink
(a) '[He went down there] and drowned.'
(b) 'He went down and [there he drowned].' Conversation_01AN_006

In the translations in (a), the PP is interpreted as the goal of the initial motion verb; in the translations in (b), it is rendered as the location of the second predicate. This ambiguity does not pose any problem in terms of the success of the communicative event, as the goal of the first verb and the location of the second verb are identical in all of these examples. Pragmatically, the two readings are compatible and there is no risk of misunderstanding. The discrepancy between the two translation options does however pose a problem for the morpho-syntactic analysis of the sentences and for the question of where the PPs structurally belong.

Text-only databases, i.e., transcriptions of recordings without associated searchable audio, do not sufficiently equip us to investigate such cases of apparent structural ambiguity. The fact is that transcription is not a pure and neutral rendering of the spoken data into text but a form of analysis. In her paper "Transcription as theory," Ochs (1979:44) stated, "transcription is a selective process reflecting theoretical goals and definitions." Clearly, every transcription filters out some features of the speech event; certain features may be omitted in transcribing; others may be added. We tend to regularize pronunciations through writing and even omit or add whole words that should or should not be there, according to our transcription standards and the current state of our morpho-syntactic analysis of the language.

An accurate traditional transcription tends to have at least some notation of intonational features through the use of different types of punctuation or, better, through a notation convention for intonation independent of punctuation. But the question is whether all intonational features marked by, say, a comma or a period are really the same. Even an elaborate notation convention for intonation would still constitute a form of interpretation.[4] Most modern transcriptions also indicate at least some pauses. But there has to be some cut-off point, and the question remains how long a pause has to be to make it into the transcript.

As Ochs described, our transcriptions tend to be tailored to what we want to do with our data, and they may, in theory, be perfectly suitable for this task. The problem is that our goals and interests may change over time and that this is a very narrow-minded and somewhat egocentric approach to a database. We do not know now what we may want to investigate in "our" data in ten years time or what other researchers, future generations of speakers, or their descendants may want to do with the data.

A database with text-audio linkage allows access to the original data, and it allows us to investigate features like pauses and intonation patterns directly (and to double-check the presence or absence of morphological material, notations of false starts, etc.). This proved to be relevant for the research questions on Saliba-Logea PPs and their position in the clause. An earlier attempt at answering this question based on a smaller text-only database showed that for some of the sequences of VERB PP VERB that do not include conjunctions or particles, the clause boundaries could not easily be determined. It is a reasonable assumption that these examples, which appear structurally ambiguous, are in effect not ambiguous at all and that the clause boundaries are clear in the audio data through intonation and pauses. We can assume therefore that ambiguity of VERB PP VERB is in the transcription rather than in the actual data, and that access to the audio will disambiguate these examples. This makes text-audio linkage of out data crucial for the morpho-syntactic analysis.

The assumption that access to the audio will resolve syntactic ambiguity certainly holds true for some of the examples, such as in the case of (14) above where the PP is preceded by a pause, which in (14') is marked by a hash (#).

(14')	*Se*	*bui-gabae*	*se*	*dobima #*	*Rossel*	*unai*	*se*	*laoma*	*snakepasis*.
	3PL	go.down	3PL	come.down	Rossel	PP.SG	3PL	come	snake.passage

'They turned away, [went down] and at Rossel Island they came to the Snake Passage.' Torres_01AC_0088-89

There was however an unexpected finding: even with text-audio linkage and therefore with direct access to the audio, some of the examples remain unclear. It turns out that PPs can be preceded and followed by a verb (i.e., a minimal clause) within the same intonation unit. Such examples of VERB PP VERB are possibly better described as vague (PP associated with both clauses) rather than as structurally ambiguous (PP associated with one clause but

[4] Transcriptions with detailed notations for intonation and pauses would be more time-consuming to create than establishing text-audio linkage. They would have the advantage, however, that noted intonational features are searchable, but obviously we need to strike a balance between as complete an annotaion as possible and the time we need to invest in order to achieve it.

it is structurally not obvious with which one) as it seems that both verbs try to claim the PP and assign it a semantic role.

This pattern seems to be restricted to cases where the first verb can take a goal argument (e.g., motion verbs: verbs like 'enter', 'throw', 'push'). The second verb does not seem to be clearly restricted; the slot may be open to any verb that expresses an activity. In such examples, it appears that the PP expresses two roles simultaneously: the goal of first and the location of second verb. It is possible that such clause sequences with vague PPs constitute a special morpho-syntactic construction (e.g., core-layer serialization). Some examples are presented in (15′) to (18′), repeated from (15) to (18) above but with the linked audio. There is no obvious clue in the intonation or the distribution of pauses that would clarify to which clause the PPs belong. Translations that include both a goal and a location PP can capture this to some extent (although, of course, they sound clumsy):

(15')	*Se*	*lao-ma*	*numa*	*wa*	*unai*	*se*	*gwau*.
	3PL	go-hither	house	ANA	PP.SG	3PL	heap

'They come to the house and gather (things) there.' Giyahi_01AA_073

(16')	*Ye*	*lao*	*unai*	*ye*	*keno-keno*.
	3SG	go	PP.SG	3SG	RED-sleep

'He went there and was resting there.' TBLaki_01AG_221

(17')	*Se*	*lao*	*nukula*	*ne*	*unai*	*maina*	*se*	*boli*	*ye*	*gehe*.
	3PL	go	jungle	DET	PP.SG	string	3PL	cut	3SG	finish

'They went to the bush and cut string there.' MakingSagoRoof_01AW_015

(18')	*Ye*	*dobi*	*unai*	*ye*	*yoli*.
	3SG	go.down	PP.SG	3SG	sink

'He went down there and drowned there.' Conversation_01AN_004

A detailed analysis of such constructions is not the focus of the paper and will be conducted elsewhere. A point of relevance for the current paper is that such vague examples need to be distinguished from examples which only seem ambiguous in text-only transcriptions.

2.2 RELEVANT DATA TREATMENT. The data treatment that enables us to disambiguate structurally ambiguous examples of VERB PP VERB and that allows us to distinguish between vague and ambiguous examples includes several components. Creating text-audio linkage as part of the regular workflow enabled us to listen to and assess the original audio recording, which was essential. Using the concordance function in Toolbox allowed us to

search and sort for relevant instances of *unai*. We used a setup of Toolbox that can play audio of single records, and we manipulated Toolbox to allow jumps from the concordance to the actual units in the respective texts and then listen to the audio. (For details on this manipulation see Andrew Margetts this issue.)

3. EXAMPLE TWO: PLURAL MARKING. In Oceanic languages, number is typically not an inflectional category of nouns, and an unmarked noun can have singular or non-singular reference. Often, only a subclass of human nouns, such as kinship terms, takes number marking and commonly the marking is not obligatory. The types of strategies found to express plural marking on nouns include reduplication, root modification with lengthening of vowels, and affixation (Lynch et al. 2002:37–39). In Saliba-Logea there is a hierarchy in which human nouns are obligatorily marked for number by a suffix, as in (19) and (20), but other nouns are not.

(19)	*wawaya*	*wawaya-o*
	child	child-PL
	'child'	'children'

(20)	*yo-gu*	*saeya*	*yo-gu*	*saeya-o*
	CLF-1SG.POSS	sibling	CLF-1SG.POSS	sibling-PL
	'my sister'		'my sisters'	

Some human nouns both reduplicate and take the suffix, as in (21) and (22):

(21)	*hasala*	*hasa-hasala-o*
	young.woman	RED-young.woman-PL
	'young woman'	'young women'

(22)	*tanowaga*	*tano-tanowaga-o*
	boss/owner	RED-boss/owner-PL
	'owner'	'owners'

If the noun takes a possessive suffix, the plural suffix attaches to the possessive marker and appears at the end of the word, as in (23) and (24):

(23)	*natu-gu*	*natu-gu-wao*
	child-1SG.POSS	child-1SG.POSS-PL
	'my child'	'my children'

(24)	*tubu-di*	*tubu-di-yao*
	granny-3PL.POSS	granny-3PL.POSS-PL
	'their granny'	'their grannies'

Nouns with non-human referents are not generally marked for number, but on the phrase level number is marked if the noun is followed by lexical modifiers or quantifiers that take a singular or plural associative suffix depending on the number of the head noun, as shown in (25):

(25)	*lulu*	*posiposi-na*	*lulu*	*posiposi-di*
	shirt	white-3SG.ASSOC	shirt	white-3PL.ASSOC
	'white shirt'		'white shirts'	

When human nouns with plural referents occur with modifiers, the noun carries the plural suffix and the modifier carries the plural associative suffix, as in (26):

(26)	*wawaya*	*gagili-na*	*wawaya-o*	*gagili-di*
	child	small-3SG.POSS	child-PL	small-3PL.POSS
	'small child'		'small children'	

The examples so far illustrate what can be termed the standard use of the plural suffix in Saliba-Logea, and similar features are attested in other Papuan Tip cluster languages. However, the older generations of speakers, at least of the Saliba dialect, comment that younger speakers occasionally use the plural suffix in context where it sounds ungrammatical. And indeed, our database includes examples of what I will call extended uses of the plural marker that do not follow the rules of affixation of the marker as outlined above.

3.1 THE NATURE OF EXTENDED USES AND WHO DRIVES THEM. A research question of interest for the morpho-syntactic description and the documentation of the language concerns the nature of the extended uses of the plural suffix and the question of who drives them. There seems to be ongoing grammatical change affecting the selection restrictions of the plural suffix, causing its spread into novel contexts.

The reported statements by older speakers suggest that it is the younger generations who drive this change. We decided to search the database for examples of extended uses of the plural marker and see (a) how they differed from the canonical use and (b) which

speakers had produced them. The findings of this search and the analysis of the data are discussed in Margetts 2008b.

The search showed that there are at least two types of extension to the standard use of the plural suffix: one is semantic in nature, the other structural. The semantic extension is evident from examples of the plural suffix on nouns with non-human referents, as in (27) and (28).

(27)	*sitowa-di-**yao***	standard use:	*sitowa*
	store-3PL.POSS-PL		store
	'stores' Notes07		'store/stores'
(28)	*beya-di-**yao***	standard use:	*beya-di*
	ears-3PL.POSS-PL		ear-3PL.POSS
	'their ears' FrogStory_01AW_095		'their ear(s)'

The structural extension is exemplified by examples of the plural suffix attaching to words other than nouns, as in (29), where it marks a possessive classifier, and in (30), where it attaches to a lexical modifier.

(29)	*yo-gu-wao*	standard use:	*yo-gu*
	CLF-1SG.POSS-PL		CLF-1SG.POSS
	'mine (pl)' notes96		'my/mine (sg/pl)'

(30)	*yama*	*gagili-di-**yao***	standard use:	*yama*	*gagili-di*
	fish	small-3PL.POSS-PL		fish	small-3PL.POSS
	'small fish' Fishing_01BQ_029			'small fish (pl)'	

Contrary to our expectations, we found that there are examples of extended uses by younger and older speakers. There was no indication that the examples came from speakers of a particular age group or that younger speakers produced more of the examples than older speakers. (But it is possible that certain types of examples are particular to younger speakers. We could not establish this because certain types of extended use, e.g., plural suffixes on possessive classifiers, are not common enough in the database.)

An unexpected finding was the fact that extended use examples were almost exclusively produced by speakers of the Saliba dialect. Logea speakers produced hardly any extended use examples, while they produced 40% of all intonation units in the corpus. In fact, Logea speakers commented that certain extended uses of the plural suffix sounded like the Saliba dialect to them. This finding is interesting, in light of the fact that there are virtually no other grammatical differences between the dialects.

3.2 RELEVANT DATA TREATMENT. To investigate the extended uses of the Saliba-Logea plural suffix and who drives them, we needed access to and searchability of both text data and metadata. We manipulated Toolbox so as to show the speaker ID in the concordance (to enable sorting by speaker), and to allow jumps from the concordance to the individual Toolbox records (for details see Andrew Margetts this issue). Jumping to the records enabled us to check the preceding and following context (necessary as some examples go over more than one intonation unit) and to verify example through the linked audio. At the time we were not yet equipped to search directly for speaker's age and dialect. We extracted information on these features from a separate database via the speaker ID.

4. CONCLUSION. In this paper I discussed how the way we treat our data can affect our analysis of basic linguistic questions. I presented some of the aspects of the data processing and treatment within a DoBeS language documentation project, including the creation of text-audio linkage and searchable metadata.

The example of the positioning of Saliba-Logea PPs demonstrates that text-audio linkage can be a prerequisite for analyzing syntactic constructions. It is the direct access to the audio of a text unit that allows us to identify and analyze constructions which may be ambiguous in the transcription but are actually straightforward if pauses and intonation are taken into account.

Even more important, audio access allowed us to identify cases where the association of the PP with either the preceding or the following clause could not be resolved through listening to the audio segment. I suggest that these instances constitute different constructions, which need to be treated separately from cases of syntactic ambiguity. The difference between these two constructions would have been missed had the analysis relied on transcripts alone. This is an important reminder that transcription is not a neutral representation of a speech event but always involves choices and filtering of what is taken to be important. The problem is that what we deem to be important in a transcription can change over time, with our knowledge of the language and with our research focus. Text-audio linkage provides direct access to the basic data without this filter, and it allows us to update transcriptions at any time.

The example of the spreading plural suffix demonstrates the importance of searchable metadata. The reasonable hypothesis that extended uses are driven by younger speakers did not hold. Instead, a totally unexpected finding came to light when examples where sorted by speaker ID, presenting what may be the first emerging grammatical difference between the Saliba and the Logea dialects. The ability to extract metadata relating to text segments—for example, information about the age and dialect of speaker—allows us to research questions relating to the linguistic behavior of different groups of speakers such as dialectal differences and ongoing language change.

Without the data treatment discussed here, some parts of the analysis would have been impossible. Or worse still, an analysis conducted on transcription data only and without access to metadata could have led to the wrong conclusions.

References

Bradshaw, Joel. 1982. Word order change in Papua New Guinea Austronesian languages. Unpublished PhD dissertation, University of Hawai'i.

Lynch, John, Malcolm Ross, and Terry Crowley. 2002. *The Oceanic languages*. London: Curzon Press.

Margetts, Andrew. 2009. Using Toolbox with media files. *Language Documentation and Conservation* 3(1): 51-86.

Margetts, Anna. 1999. *Valence and transitivity in Saliba, an Oceanic language of Papua New Guinea*. MPI Series in Psycholinguistics. Nijmegen: Max Planck Institute for Psycholinguistics.

Margetts, Anna. 2004. Core-layer junctures in Saliba. In *Complex verbs and serialization in Oceanic languages*, ed. by Isabelle Brill and Françoise Ozanne-Rivierre, 69–89. Empirical Approaches to Language Typology. London, New York: Mouton de Gruiter.

Margetts, Anna. 2005. Positional slots in Saliba complex verbs. *Oceanic Linguistics* 44(1):65–89.

Margetts, Anna. 2008a. Transitivity discord in some Oceanic languages. *Oceanic Linguistics* 47(1):31–44.

Margetts, Anna. 2008b. Spread of the Saliba-Logea plural marker. Ms.

Ochs, Elinore. 1979. Transcription as theory. In *Developmental pragmatics*, ed. by Elinor Ochs and Bambi B. Schieffelin, 43–72. New York: Academic Press.

Oetzel, Rainer, and Sabine Oetzel. 1997. Orthography and phonology description of Saliba. Ms. Ukarumpa: SIL.

Ross, Malcolm. 1988. *Proto Oceanic and the Austronesian languages of Western Melanesia*. C-98. Canberra: Pacific Linguistics.

[1] DoBeS, Documentation of endangered languages. http://www.mpi.nl/DOBES/

[2] Transcriber to Toolbox converter. http://linguisticsoftwareconverters.zong.mine.nu/

Anna Margetts
anna.margetts@arts.monash.edu.au